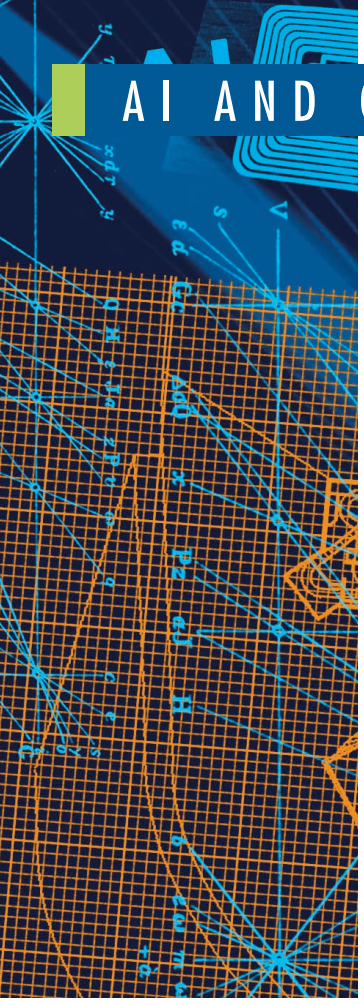
Knowledge-Based Linguistic Annotation of Digital Cultural Heritage Collections

Ruotsalo, T.; Aroyo, L.M.; Schreiber, A.Th.

published in
IEEE Intelligent Systems
2009

DOI (link to publisher)
10.1109/mis.2009.32

document version
Publisher's PDF, also known as Version of record



Knowledge-Based Linguistic Annotation of Digital Cultural Heritage Collections

Tuukka Ruotsalo, *Helsinki University of Technology*

Lora Aroyo and Guus Schreiber, *Vrije Universiteit Amsterdam*

A method for automatically annotating objects in digital cultural heritage collections uses structured vocabulary concepts and their metadata schema roles.

A growing number of cultural heritage collections are available in digital form. Although techniques exist to physically preserve digital objects, annotating and searching such collections is far from trivial. The main annotation approaches are based on text and structured vocabularies. Text-based retrieval can use textual descriptions or a set of keywords accompanying an image, but this approach limits the range of successful queries to the indexer's interpretation or to the individual words appearing in the textual descriptions.[1] For example, in a traditional text-based search, a query to "retrieve all paintings" wouldn't return panels or portraits, although a user might be interested in them because they're special cases of paintings. Nor would a traditional text-based search distinguish between, for example, Paris as place where a painting was produced and Paris as a place depicted in the painting.

Recent research has replaced keyword-based annotation with annotation using structured vocabularies and schema-based metadata that explicate the concepts (for example, Paris as a city) and roles (for example, Paris as a subject matter). Such lightweight semantic background knowledge has enhanced the performance of retrieval methods, especially in applications that require highly structured queries.[2,3] Applications that use a structured vocabulary to help searchers navigate can benefit from such data.[4]

Manual annotation is predominant in the cultural heritage domain.[5] This can be a tedious process that leaves many objects stored with incomplete annotations at best and no annotation at worst. Typically, objects in cultural heritage collections are accompanied by a textual description. However, traditional information extraction isn't completely suitable for automatic annotation, which requires both concept identification (mapping word occurrences or word chunks to concept instances in structured vocabularies—Paris as a city) and role identification (identifying annotation schema roles that these instances play in the text—Paris as a subject matter).

In this article, we present a method that uses natural language processing techniques and background knowledge in the form of structured vocabularies to automatically identify concepts and their roles from text descriptions. Many current methods perform relatively well in identifying concepts,[6] so we focus on role identification. Recent work in role identification has aimed at predicate-argument structure identification.[7] We don't consider predicate-argument structures, but instead use annotation schema roles as the target role set for the method. Furthermore,

1541-1672/09/$25.00 © 2009 IEEE
Published by the IEEE Computer Society

our focus is on constituents either that are named entities or for which correspondences can be found from structured vocabularies. We include an empirical evaluation of the automatic method that achieves performance close to the level of human annotators.

Structured Vocabularies and Metadata Schema Roles

Several structured vocabularies are available in the cultural heritage domain. The vocabularies provide a controlled set of concepts and instances to be used in annotation. For this study, we used three structured vocabularies from the J. Paul Getty Trust (see www.getty.edu/research/conducting_research/vocabularies for more information on licensing for research and regular uses of the Getty vocabularies) to cover the subdomains of persons, organizations, geographical locations, and terms specific to cultural heritage, and the WordNet lexical database to cover common lexical terms:

- The *Art and Architecture Thesaurus* (AAT) is a structured vocabulary of around 34,000 concepts, including 131,000 terms, descriptions, and other information relating to fine art, architecture, decorative arts, archival materials, and material culture.
- The *Getty Thesaurus of Geographic Names* (TGN) is a structured vocabulary containing around 912,000 records, including 1.1 million names, place types, coordinates, and descriptive notes, focusing on places important for the study of art and architecture.
- The *Union List of Artist Names* (ULAN) is a structured vocabulary containing around 120,000 records, including 293,000 names and biographical and bibliographic information about artists and architects, including a wealth of variant names, pseudonyms, and language variants.
- *WordNet* (WN) is a general lexical database that organizes nouns, verbs, adjectives, and adverbs into synonym sets, each representing one underlying lexical concept. WordNet also provides relations for hyponymy, meronymy, and troponymy. (WordNet is available in RDF/OWL format from www.w3.org/TR/wordnet-rdf.)

The concepts in the structured vocabularies are typically ordered in subsumption or meronymical hierarchies.

> Metadata schema roles used in this study enable the annotation of the most important artwork features and its subject matter.

For example, "canvas" *is-a* "material," or "Amsterdam" is a *part-of* "The Netherlands." The vocabularies also provide a set of spelling variants or synonyms for each concept.

In the annotation of cultural heritage objects, factual information is typically distinguished from the subject matter depicted. A metadata schema enables this further structuring of the annotation.[5] The schema consists of a set of roles that indicate how vocabulary concepts are linked to the artwork. For example, a person's name might appear in the artwork's *title*, or as its *creator*, or in a *subject people* role. For this study, we used a Visual Resources Association (VRA, www.vraweb.org/resources/datastandards/vracore3/index.html) specialization of the Dublin Core metadata schema tailored to the needs of artwork annotation. (An unofficial OWL specification of the VRA 3.0 elements, including links to the Dublin Core, is available at http://e-culture.multimediaitn.nl/resources.)

Cataloguing experts at Rijksmuseum Amsterdam defined the most important metadata schema roles used in this study, as shown in Table 1 (see next page). These roles enable the annotation of the most important artwork features—for example, who created it, where, and when. The metadata schema also contains roles to indicate the artwork's subject matter—for example, who, what, where, and which time period it depicts. This schema makes it possible to express, for example, that an artwork's *creation location* is "Amsterdam," its *material* is "canvas," or its *subject location* is "Amsterdam." The metadata schema also defines a value range for each role, determining the vocabulary or vocabulary branch from which to select the role values.

Linguistic Annotation

Because objects in digital cultural heritage collections are typically accompanied by natural language descriptions, the annotation process can benefit from extracting metadata automatically from these descriptions. However, alternative syntactic and lexical realizations of semantic arguments are widespread.

For example, consider the following sentences:

- The work was created in Arles in 1888 by Van Gogh.
- In Arles, Van Gogh painted the still life in 1888.

Both sentences express the same semantic content but have different syntactic and lexical realizations; for example, the hyponymous verbs *create* and *paint* refer to the same underlying event, and the hyponymous nouns *still life* and *work* refer to the same artwork. The

Table 1. Dublin Core Visual Resources Association metadata schema roles selected for annotation.

Role	Explanation	Value range*
Work type	Specific type of artwork being described.	AAT, WN
Title	Title or identifying phrase given to an artwork.	Literal
Material	Substance of which an artwork is composed.	AAT, WN
Technique	Production or manufacturing processes, techniques, and methods incorporated in the fabrication or alteration of the artwork.	AAT, WN
Creator	Names, appellations, or other identifiers assigned to an individual, group, or corporate body that has contributed to the design, creation, production, manufacture, or alteration of the artwork.	ULAN, Literal
Creation date	Date or range of dates associated with the creation, design, or production of the artwork.	Literal
Repository location	Geographic location and/or name of the repository locations entity whose boundaries include the artwork.	Literal
Creation location	Geographic location and/or name of the creation locations entity whose boundaries include the artwork.	TGN
Style period	A defined style, historical period, group, school, dynasty, movement, etc., whose characteristics are represented in the artwork.	AAT, WN
Cultural context	Name of the culture, people (ethnonym), or adjectival form of a country name from which an image originates, or the cultural context with which the artwork has been associated.	AAT, WN
Subject term	Terms or phrases that describe, identify, or interpret the artwork and what it depicts or expresses. These include generic terms that describe the work and the elements that it comprises.	AAT, WN
Subject people	Terms or phrases that describe, identify, or interpret particular people.	ULAN, WN, Literal
Subject location	Terms or phrases that describe, identify, or interpret geographic places.	TGN, WN
Subject date	Terms or phrases that describe, identify, or interpret time.	Literal

*Acronyms: AAT = Art and Architecture Thesaurus, TGN = Thesaurus of Geographic Names, ULAN = Union of Artist Names, WN = WordNet.

positions and grammatical functions of the sentence constituents and voice vary; for example, the first sentence is in passive voice and the second in active voice. In sentences written in passive voice, the subject receives the action expressed in the verb—that is, the subject is acted upon. In sentences written in active voice, the subject performs the action expressed in the verb—the subject acts. In other words, the constituents' grammatical functions can vary, while their semantic roles are the same.

In addition, named entities and concept chunks that represent intuitive subdivisions of a sentence are important; for example, it was "Van Gogh" who painted the "still life" and not "Van" and "Gogh" or "still" and "life."

We use a technique known as *semantic role labeling*,[7] where the sentence's syntactic features and predicate are used to predict each constituent's role. Because we perform the annotation using structured vocabularies, we focus on constituents that are named entities or that have a concept corresponding to these vocabularies.

Figure 1 presents the overall architecture of our approach. It consists of three phases: linguistic analysis, concept identification, and role identification. We first perform the linguistic analysis for a sentence in the textual description, then use the resulting syntactic features to perform the concept identification. Finally, we perform the role identification over the result of both the linguistic analysis and the concept identification. Concept identification determines the concepts that have correspondences in the vocabularies and are therefore candidates for annotation. Role identification determines the semantic role, if any, that these concepts play in the annotation.

Phase 1: Linguistic Analysis

Linguistic analysis is a process that provides information about a natural language sentence's syntactic features. It consists of four steps:

- *Named-entity tagging* is an information-extraction task that locates and classifies atomic text elements into predefined categories. In this study, we used a named-entity recognition system to produce the following classes: persons, organizations, locations, and miscellaneous named entities.
- *Part-of-speech (PoS) tagging* determines the correct syntactic class (a part of speech, such as a noun or verb) for a particular word given its current context in the sentence. PoS tagging involves disambiguation between multiple part-of-speech tags.
- *Morphological analysis* addresses the inflectional and compounding processes in word formation to determine inflectional properties—for example, gender (male, female, or neuter), number (plural or singular), and case (nominative, accusative, or dative). Together with PoS information, this process delivers a word's morphosyntactic properties.

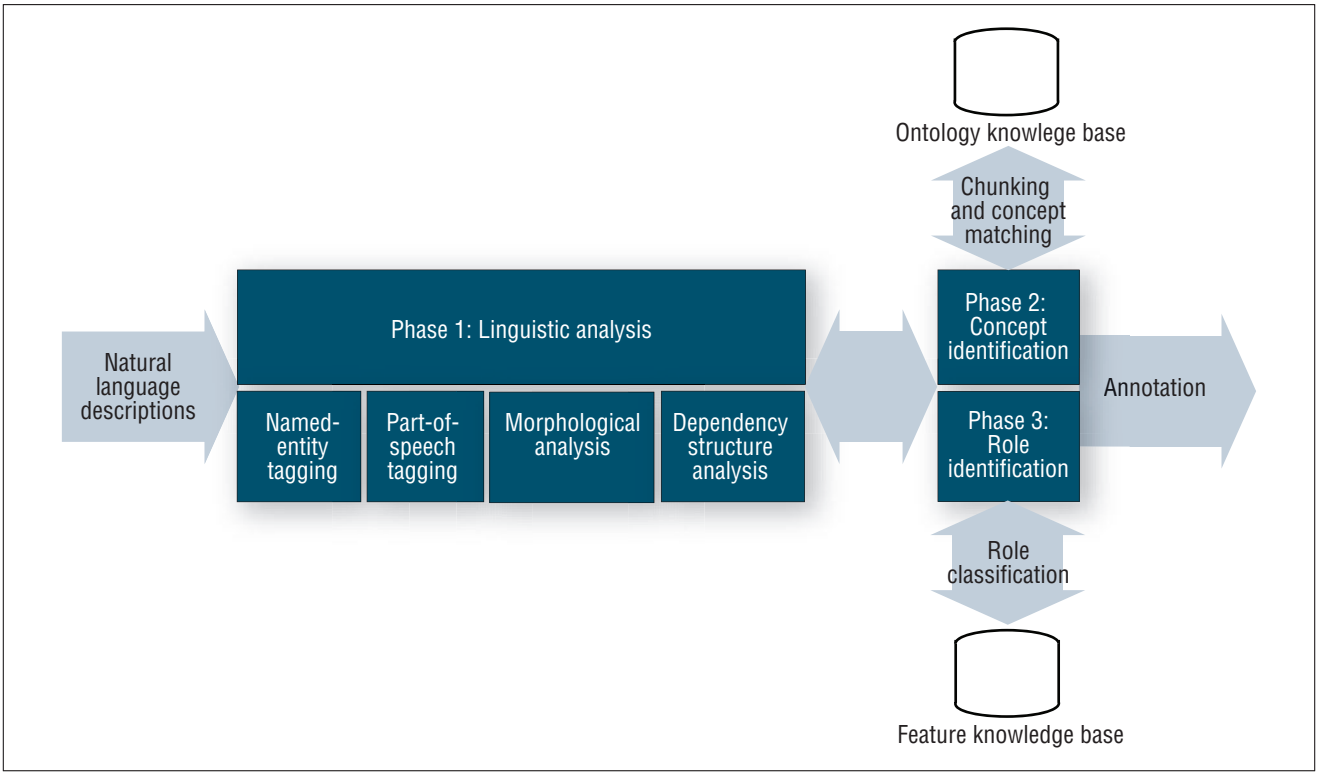


Figure 1. System architecture. In three phases, our approach to annotation performs linguistic analysis, concept identification, and role identification.

- *Dependency structure analysis* consists of analyzing two or more linguistic units that dominate each other in a syntax tree. The main outcome of dependency structure analysis is the sentence's internal dependency structure. It consists of grammatical functions, such as subject and direct object, that help identify participants of the sentence's events or verb phrases.

As Figure 1 illustrates, both the concept and role identification phases use the linguistic analysis. For example, the morphological analysis provides lemmas for words and identifies date or number chunks in the text. PoS tagging is used to separate verbs, adjectives, and nouns. For instance, the word "works" can be either a verb or a plural noun, depending on the sentence context. In concept identification, we use this information to disambiguate the vocabulary concept correspondences. In role identification, we can use the dependency paths from the dependency structure analysis to identify roles in complex sentence structures; for example, verbs are often internally dependent on other verbs in a sentence. In the sentence, "This painting is believed to be painted in Amsterdam," the main verb is "painted," but the sentence also has two other verbs—namely, "is" and "believed." Using the dependency structure analysis, we can determine the relations between these verbs and identify the main verb—in this case, "painted."

Similar dependency structures are also useful in identifying paths to a sentence's nouns. In the sentence, "The painting was painted in 1888 in Arles," the parser gives the dependency path *prep-in* (denoting the preposition "in" between the verb and the constituent) to connect both "1888" and "Arles" to the verb "painted." This information can be used to enhance role identification.

For linguistic analysis, we used the Stanford Lexical PCFG (Probabilistic Context-Free Grammar) Parser,[8] trained on the English Penn Treebank.[9] The parser performs the PoS tagging and the dependency structure analysis; we set it to return only the parse tree with the highest confidence level.

Phase 2: Concept Identification

Concept identification is the process of defining meaningful units from a sentence and mapping these to concepts in the structured vocabularies. It consists of two steps:

- *Chunking* identifies meaningful units, such as named entities, noun phrases, or verbal groups.
- *Concept identification* compares the chunks with a set of vocabularies.

Chunking is performed in four steps. First, all named entities are considered as chunks to be included for further analysis, even if they don't occur in the structured vocabularies. Second, any text appearing in quotes is chunked to a single phrase. Third, because a temporal vocabulary containing

Table 2. Linguistic and vocabulary-based features used in role identification.

Feature	Explanation	Example in Figure 2
Verb identifier	The identifier for a verb concept in a structured vocabulary.	The verbs "present" and "portray" are normalized to the verb identifier *portray* sense 4 from WordNet.
Voice	The active or passive voice in a sentence verb.	The verb "portray" is in passive voice.
Position	The constituent's position appears before or after the verb.	The constituent "regalia" occurs after the verb.
Constituent identifier	The constituent's identifier in the vocabulary.	The constituent "regalia" has an identifier from AAT and WordNet.
Constituent PoS	The constituent's PoS tag.	The constituent "regalia" is tagged as singular or mass noun (NN).
Partial PoS path	The partial PoS path in the parse tree from the parse constituent to the verb or predicate. The verb and the constituent are not included in the partial path.	The partial PoS path for the constituent "regalia" contains only the preposition or subordinating conjunction "IN."
Partial dependency path	The dependency tags of the constituents on a path through the parse tree.	The partial dependency path for "regalia" contains only the *[prep-in]* dependency tag.
Constituent vocabulary base	The identifier of a vocabulary where the concept correspondence was found during the concept identification phase.	The vocabulary bases for the constituent "regalia" are WordNet and AAT.
Vocabulary root	The constituent's vocabulary-root concept.	The vocabulary roots for the constituent "regalia" are the "artifact" concept from WordNet and the "object genres" and "object groupings and systems" concepts from AAT.
Constituent word type	One of the following: number, person, organization, place, miscellaneous, or noun.	The constituent "regalia" has the word type "noun."

correspondences to temporal expressions wasn't available for this study, the chunking method tags temporal expressions on the basis of temporal cue words, such as "13th century," then transforms them into a four-digit number format. We found that these general rule patterns covered the cases appearing in the selected data set.

In the final step, we first match the rest of the text as bi-word-grams against the structured vocabularies. In cases where we found no correspondence for the bi-word-gram, we matched a single word against the structured vocabularies. For example, bi-word-grams such as "young woman" or "in fact" are chunked, because the vocabularies have corresponding concepts for these chunks. However, a bi-word-gram such as "green vase" is not chunked, because the vocabularies include no correspondence, so the bi-word-gram isn't considered useful to further annotation. In other words, the chunk "young woman" can be an important concept in the art domain, while "vase" is a type of object independent of the color property that it possesses. In this process, only the noun phrases are matched against the WordNet noun facet and AAT. The chunks for which correspondence is found are passed further to the concept identification phase.

Concept identification is performed differently for noun phrases, verbs, and named entities. We match the named entities against the vocabularies. Those tagged with a *person* or *organization* are first matched in Getty's ULAN; if no correspondence is found, the lookup continues in other vocabularies. This is necessary because fictional characters often appear as subjects of artwork (for example, "Venus"). The named-entity tagger will tag them as *person* or *organization*, and only WordNet will later properly identify them. Our method matches the person names first on the basis of common substrings separated by white space and then on the basis of edit-distance between uppercase letters. The method can thus match spelling variants with abbreviations that aren't directly in ULAN. For example, the name "J.M.W. Turner" is matched to "Joseph Mallard William Turner." The words tagged as *miscellaneous* are first looked up in ULAN, then in TGN, AAT, and finally WordNet.

We match noun phrases and verbs against AAT and WordNet. With AAT, we usually find only one matching concept. On the other hand, WordNet contains a number of word senses, so we select the most typical sense of the word according to its usage rank. WordNet also contains expressions for verbs. Here, we use the PoS tag to distinguish verbs and nouns that have the same syntactic form. We also use this information in word stemming. First, we search exact matches; if we find no exact match, we use a word form. We pass the named entities to further analysis even if we find no correspondence in vocabularies. Other chunks are considered relevant only if we find correspondence in the vocabularies.

We performed named-entity recognition using the Stanford Named Entity Recognition System,[10] trained on the CoNLL-2003 English training data.[11] For word stemming, we used the Snowball stemmer (http://snow-

Figure 2. Example of collapsed typed dependencies and role-identification features for the constituent "regalia" used in role identification.

Original sentence:

The queen is portrayed in her full regalia.

Collapsed typed dependencies:

det(queen-2, The-1)
nsubjpass(portrayed-4, queen-2)
auxpass(portrayed-4, is-3)
poss(regalia-8, her-6)
amod(regalia-8, full-7)
prep_in(portrayed-4, regalia-8)

Features for the constituent "regalia":

1. Verb identifier: http://www.w3.org/2006/03/wn/wn20/instances/synset-portray-verb-4
2. Passive voice: true
3. Position before verb: false
4. Constituent identifier: http://www.w3.org/2006/03/wn/wn20/instances/synset-regalia-noun-1, http://e-culture.multimedian.nl/ns/getty/aat#300185696
5. Constituent PoS: NN
6. Partial PoS path: IN
7. Partial Dependency Path: prep-in
8. Constituent Ontology Base: WN, AAT
9. Ontology root: http://www.w3.org/2006/03/wn/wn20/instances/synset-artifact-noun-1, http://e-culture.multimedian.nl/ns/getty/aat#30018711, http://e-culture.multimedian.nl/ns/getty/aat#300264092
10. Constituent Word Type: Noun

ball.tartarus.org/algorithms/porter/stemmer.html) for English, enhanced with a list of irregular verb forms. For concept identification, we used the Multimedia e-Culture API[4] extended with the edit distance for uppercase letters in the case of person names.

Phase 3: Role Identification

Many current systems perform relatively well in concept identification, but they often fail in role identification.[6] For example, for the word chunk "Van Gogh," existing techniques can determine that the entity is an instance of the concept "Person." However, determining that the same entity plays the role of "Creator" of a given artwork is more difficult.

The difference between concept and role identification is that the concept types of the instances are the same independently of the context they're used in, and the roles in an annotation can vary independently of the concept type. For example, the word "Rembrandt" in the art domain will generally be typed as "Person." However, in the role-identification task, "Rembrandt" can take various roles, such as *creator* or *subject* of an artwork.

To assess the role for the determined concepts, we built a separate classifier using two types of features: syntactic (produced by linguistic analysis) and semantic (derived from the vocabularies on the basis of concept identification). The syntactic features we use are a subset of the features presented by Daniel Gildea and Daniel Jurafsky[7] and by Sameer Pradhan and his colleagues.[12] Table 2 presents the features used in the classifier, and Figure 2 shows examples of the feature instantiations for the constituent "regalia."

Typically, the use of synonyms, hyponyms, or phrasal expressions vary the linguistic realization of natural language sentences. This variation causes sparsity of the data that can be reduced using synonym sets and hyponymy hierarchies available in the vocabularies. To overcome the synonym problem, verb identifier and constituent identifier features are normalized to vocabulary concepts in the concept identification phase. For example, in Figure 2, the verbs "present" and "portray" are normalized to the verb identifier *portray* sense 4 in WordNet. To overcome the hyponym problem, a reasoning procedure infers the vocabulary-root feature. For example, in Figure 2, the constituent "regalia" is inferred to be an "artifact," an "object genre," and an "object grouping and system." Similarly, the constituent "queen" would be inferred to be "people."

In addition to variation in linguistic realization, natural language sentences can vary in syntactic realization. Several techniques can overcome the problems caused by syntactic variation. The verb identifier together with the voice and position features give an approximation of the constituent's context in the sentence. This supports syntactic distinctions such as verb objects in active voice from verb subjects in passive voice. For example, in the sentences "Van Gogh painted the painting" and "The painting was painted by Van Gogh," this technique can determine that "Van Gogh" is the one who paints.

Path features capture the natural language patterns connecting verbs and constituents. For example, in the sentence, "The painting was painted in 1888 in Arles," the PoS path from the verb "painted" to the constituent "Arles" would be *[in]*. Such prepositional phrases increase the role identification's accuracy.[7]

In addition to the PoS path, a sentence's collapsed dependency structure (see Figure 2) is a source for a path feature. We removed the constituent itself and the predicate/verb from the path features because the information

about the verb and the constituent itself are already available as individual features. In addition, we removed all adjectives and adverbials (PennTreeBank tags under JJ, RBR, and RBS) from the partial path because they are often irrelevant for role identification. For example, in the sentence, "Van Gogh painted a beautiful painting," "painting" should be defined as a "work type" independently of whether it's beautiful. If the adjective or the adverbial is chunked in the concept identification phase, it's already merged with its referent constituent and therefore removed from the path in the concept identification phase. We also removed verb frames that the dependency structure analysis determines to be negations.

For the role-identification task, we built a classifier using support vector machines (SVMs) that have proven performance in text classification.[12] SVMs use a statistical learning algorithm that works by finding an optimal hyperplane in a feature space. The optimal hyperplane maximizes the separation between roles on the basis of the feature space. Each possible value for each feature is encoded as a Boolean nominal feature having value 1 or 0. Because the natural language and therefore our feature set is highly nonlinear, we use a polynomial kernel function. The polynomial kernel has a degree of 2; a cost-per-unit violation of the margin, $C = 1$; and tolerance of the termination criterion, $e = 0.001$. We implemented the classifier using the Weka 1.5 machine-learning toolkit (www.cs.waikato.ac.nz/ml).

Experimental Setup

The annotation of artwork objects isn't an isolated activity. It's a key element of integrated collection management. In the worldwide massive digitization of cultural heritage collections, automating the annotation process is critical. It's unrealistic to assume that human annotators can annotate these continuously fast-growing collections. Automatic linguistic annotation can produce structured annotation for collections where natural language descriptions for the objects are available.

In this study, we focused on role identification, where annotation is performed with a metadata schema. We've addressed two specific questions in the study: What accuracy does our annotation method achieve in role identification compared to human annotators? Does the usage of structured vocabularies as background knowledge increase the performance of the method?

The annotation of artwork objects isn't an isolated activity. It's a key element of integrated collection management.

Data Set

We evaluated our annotation method using the ARIA collection from Rijksmuseum Amsterdam (www.rijksmuseum.nl/aria). We randomly selected 250 artworks (for example, images of statues, miniature models, and paintings) for the experiments. All artworks were accompanied by English natural language descriptions that typically describe what the artwork depicts, what material the artwork is made of, who created it, and where and when it was manufactured. The natural language descriptions also contained additional information about the people involved, things typical to the time period, and general history related to the artworks. In other words, the task was not only to identify concepts and roles for concepts in the text but also to separate the information in the natural language descriptions that humans found relevant for annotating the artwork from the information they didn't find relevant. To enable this, we added *none* to the set of possible roles.

Evaluation Methods

We evaluated our annotation method's performance in two ways. We first compared it to a human-created gold standard to determine the method's overall accuracy. Second, we compared it to a baseline method (one without the use of structured vocabularies) to study the role of the background knowledge. Finally, to put the results in perspective, we determined the performance humans achieved in the annotation task. The performance of human annotators was also important because the quality of the machine-learning simulation results is highly dependent on the agreement and consistency of the training data. Thus, if humans have low agreement, the method can't be expected to perform consistently.

We conducted a user experiment to produce the gold standard. Fourteen computer science students and faculty members from Vrije Universiteit Amsterdam participated. All participants had previous experience with structured vocabularies, annotation tools, and metadata in the cultural heritage domain. The participants annotated a total of 250 artworks using the annotation form shown in Figure 3. The number of documents annotated by individual participants varied between 10 and 20. Three artworks were annotated multiple times by different annotators, which let us measure the interannotator agreement. We used a k-fold cross-validation, which resulted in 60 comparisons of the artworks and 2,066 comparisons of the individual concepts.

The queen is portrayed in her full regalia. Everything here emphasizes her royal status: the crown, her ermine robes and the canopy. This official portrait was made in the studio of the painter Frans Pourbus II. It is a copy of the portrait in the Louvre in Paris. The queen is Marie de Médicis (1573–1642), a member of the renowned Italian family and wife of the King of France, Henry IV (1553–1610), whom she married in 1600. Henry IV was an ally of the Dutch Republic in the struggle against Spain. In 1638 Marie de Médicis, now a widow, visited Amsterdam, where she was received with great ceremony. For the occasion Joachim van Sandrart painted a militia painting in which she is also portrayed.

Annotation form

The queen is portrayed in her full regalia .

queen — Term as a Subject Matter / Depicted Object / Depicted Event / Keyword

regalia

Everyth … **es and the canopy .**

royal

status

crown

robes

canopy

None
Term as a Subject Matter / Depicted Object / Depicted Event / Keyword
Place as Subject Matter or Keyword / Depicted Place
Person / Organization / Fictional Character as a Subject Matter, Keyword or Depicted
Time as a Subject Matter, Keyword or Depicted Time
Work Type
Title
Material
Technique
Creator
Date of Manufacturing / Date of Discovery
Manufacturing Location / Discovery Location
Repository Location
Style or Period
Cultural Context

This official portrait was made in the studio of the painter Frans Pourbus II .

Figure 3. Annotation interface. Experiment participants viewed an artwork image accompanied by a text description. Then they selected a correct role for every word chunk found in the text.

We used the same features to create the baseline method as we used in our annotation method, excepting the vocabulary-based features (vocabulary root and constituent vocabulary base). To ensure a similar amount of information available for both methods, we replaced the values of vocabulary-based features with the original word chunks for the baseline method.

We measured the performance of our annotation method, the baseline method, and the human annotators using precision, recall, and F1 measures. We calculated precision as the share of correctly classified examples out of all classified examples assessed for the measured role. Recall was the share of correctly classified examples out of all relevant examples assessed for the measured role. All relevant examples were those assessed for the role in the human-created gold standard. All classified examples of a role were those classified to a certain role by our automatic annotation method.

We conducted a simulation to test the performance of our annotation method and the baseline method. We randomly selected a set of 70 percent (n = 175) of the artworks to use as a training set and 30 percent (n = 75) to use as an evaluation set. This resulted in 8,807 individual concept occurrences in the training set and 3,985 in the evaluation set. The split into training and evaluation sets on the artwork level ensured that the method couldn't benefit from multiple occurrences of the same word chunk in the learning phase. For example, the creator's name typically occurred in multiple sentences in an artwork's natural language description. So we wanted to exclude the possibility that the method would benefit from learning the actual word chunk in the same artwork context already in the training set.

To investigate the possibility of chance in the measurements, we used Cohen's Kappa, which measures concordance between two classifiers or annotators using nominal data.[13] Kappa varies between –1.0 and 1.0. The degree of concordance is considered moderate if Kappa is larger than 0.4 and substantial if Kappa is larger than 0.60. To measure the statistical significance of the performance differences between the baseline method and the method with background knowledge in the form of structured vocabularies, we used a chi-square test by comparing the number of correctly classified examples (true positives and true negatives) and the number of wrong classified examples (false positives and false negatives). SVMs for multirole classification problems are implemented by building separate binary classifiers for each role, so we also wanted to measure the significance of the results for each role. This is why we calculated Cohen's Kappa and chi-square tests for each role separately.

In a classification task, precision and recall are vulnerable measures. We employed the F1 measure as the main evaluation metric because it combines precision and recall into a single metric and favors a balanced performance of the two metrics. In the experiment, all and only all of the concepts were classified, so precision and recall for the total data set are equal and therefore called accurate. We performed a two-tailed t-test for the F1 measures of our method and the baseline method to ensure a statistical significance of the results. Because of a rather small number of F1 measures that aren't necessarily normally

Table 3. Results of the role identification, where P = precision, R = recall, F1 = F1 measure, and K = Cohen's Kappa.

	Data		Baseline				Our method				Humans			
Role	Training set	Evaluation set	P	R	F1	K	P	R	F1	K	P	R	F1	K
Creation date	209	80	61.3	85.0	71.2	0.71	70.8	78.8	74.6	0.74	95.0	95.0	95.0	0.95
Work type	456	194	60.7	43.8	50.9	0.49	59.2	53.1	56.0	0.54	63.7	62.4	63.0	0.61
Creator*	438	180	59.7	71.7	65.2	0.63	72.1	68.9	70.5	0.69	90.9	98.0	94.3	0.94
Subject location	307	134	44.4	44.8	44.6	0.43	52.0	38.1	44.0	0.42	61.3	55.9	58.5	0.56
Material	179	76	53.3	53.3	39.7	0.39	58.0	67.1	62.2	0.61	93.3	93.3	93.3	0.93
Technique	40	21	2.3	4.8	3.1	0.02	5.7	9.5	7.1	0.07	46.2	37.5	41.4	0.41
Cultural context	189	94	43.8	48.9	46.2	0.45	54.1	42.6	47.6	0.47	51.6	45.7	48.5	0.46
Style period	40	13	6.7	7.7	7.1	0.07	4.5	7.7	5.7	0.05	66.7	40.0	50.0	0.50
Title	79	16	20.0	25.0	22.2	0.22	18.2	25.0	21.1	0.21	36.4	33.3	34.8	0.33
Subject people	841	368	51.7	45.1	48.2	0.43	49.3	56.5	52.7	0.47	75.6	81.4	78.4	0.76
Repository location	52	28	68.8	39.3	50.0	0.50	52.4	39.3	44.9	0.45	50.0	44.4	47.1	0.47
Subject Term*	2,783	1,211	56.3	73.4	63.7	0.45	61.1	70.7	65.5	0.49	63.6	67.3	65.4	0.47
Creation location	157	67	70.8	50.7	59.1	0.59	50.7	50.7	56.7	0.56	47.1	53.3	50.0	0.49
None*	2,955	1,451	66.6	53.7	59.4	0.56	68.5	59.6	63.8	0.45	64.9	62.1	63.4	0.44
Subject date	82	52	35.0	13.5	19.4	0.19	50.0	50.0	50.0	0.49	75.0	75.0	75.0	0.75
Total**	8,807	3,985	57.8	57.8	57.8	0.49	61.2	61.2	61.2	0.54	65.1	65.1	65.1	0.58

* ($p < 0.05$)
** ($p < 0.01$)

distributed, we also performed the Wilcoxon signed-rank test.

Results

Table 3 shows the results of the experiments. Our annotation method with the full set of features achieved an average accuracy of 61.2 percent (Cohen's Kappa = 0.54), and the baseline method, which used only statistical and lexical features, achieved an average accuracy of 57.8 percent (Cohen's Kappa = 0.49). The difference between our annotation method and the baseline is statistically significant ($p < 0.01$). The human annotators' accuracy was 65.1 percent (Cohen's Kappa = 0.58).

Figure 4 shows the F1 measures of our annotation method, the baseline method, and the human annotators. The overall F1 measure of our method compared to the F1 measure of the baseline method was almost statistically significant according to the two-tailed t-test ($p < 0.06$). The two-tailed t-test shows statistical significance ($p < 0.05$) when we excluded roles that had fewer than 30 evaluation examples. The Wilcoxon signed-rank test showed statistical significance ($p < 0.05$) for the whole role set. The F1 measure for humans is always higher than the F1 measure for either of the methods except for two roles (*subject* and *creation location*). A possible explanation for this error is the low number of examples in the multiple annotated part of the data set and the use of k-fold cross-validation. Cohen's Kappa shows moderate to substantial overall agreement for the human annotators.

For roles where the difference between our annotation method and the baseline method were statistically significant—specifically, *subject term*, *none*, and *creator* ($p < 0.05$)—our annotation method shows better results on the F1 measure than the baseline method does. In two roles, *subject term* and *creator*, the baseline has higher recall but substantially lower

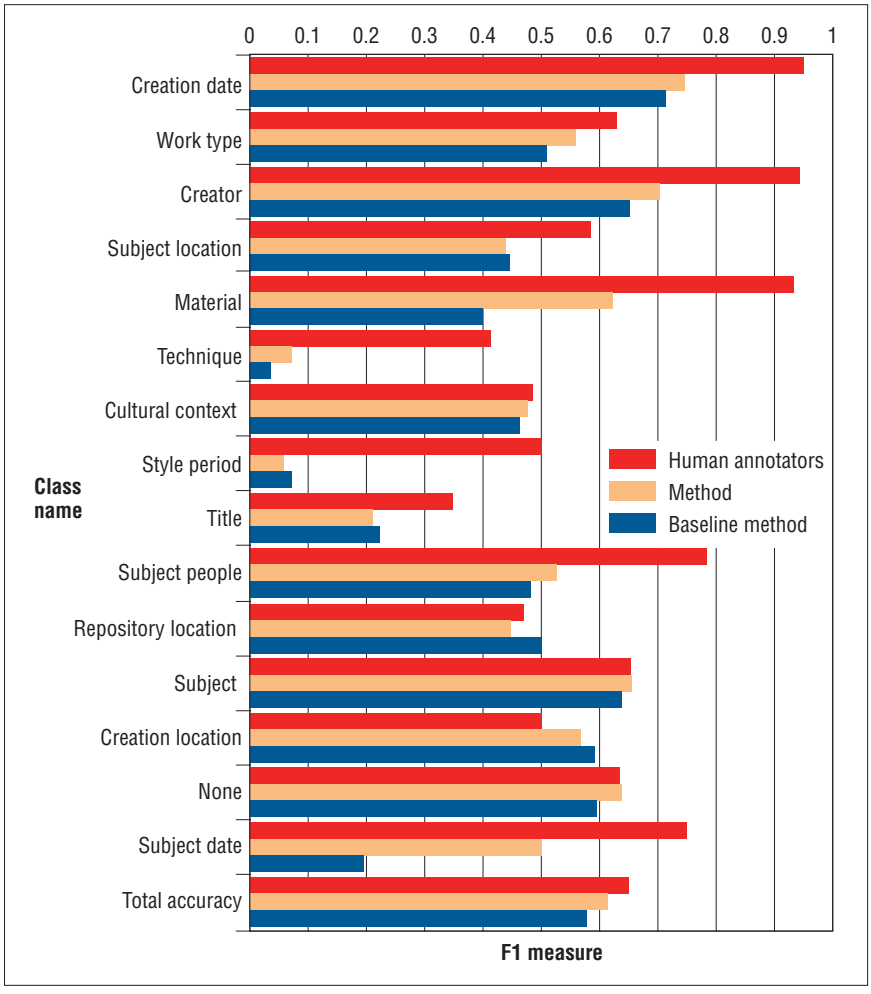


Figure 4. F1 measure of the simulation. The results show performance for human annotators, our knowledge-based method, and the baseline method in the role identification task.

precision. We used qualitative analysis to identify two main reasons for this.

First, the named-entity tagger made mistakes in the *creator* role. In some cases, our annotation method was able to correct the mistakes and make the distinction between the subject matter and the factual information on the basis of concept identification. For example, creators were more often listed in the ULAN vocabulary than were the persons depicted in the paintings. Also, some fictional characters that were commonly depicted in paintings or were the subject matter in statues, such as "Buddha," were found in WordNet. The named-entity tagger also sometimes wrongly classified geographical locations, but the concept-identification process was able to identify them.

Second, our annotation method classified the *subject* and *none* roles more accurately. On the basis of a qualitative analysis, we think this was because the machine-learning method was able to use the vocabulary-root feature to distinguish relevant from irrelevant word chunks. The use of vocabularies also seemed to enhance the distinction of the role *subject term* from the role *subject people*, such as "prince," "man," or "daughter," and from the role *work type*, such as "landscape," "drawing," or "head."

In two roles, *technique* and *style period*, performance for both the interannotator agreement and our method was very low. (recall between 7.7 percent and 9.5 percent, precision between 4.5 percent and 5.7 percent, and Cohen's Kappa between 0.05 and 0.07). Qualitative analysis revealed that, in general, these roles were used inconsistently in annotation.

AAT has a taxonomy for "processes and techniques" and "style periods," which are meant to be used as value ranges for the corresponding fields. However, our study participants often classified other concepts to possess these roles. For example, they classified "Terracotta" as *technique*, whereas the root class in AAT is "materials." The participants also classified "sculpture" as *technique*, whereas its root class is "visual works" in AAT. In classifying style periods, the annotators used temporal expressions such as "11th century," which didn't benefit from background knowledge. Structured vocabularies didn't contain some of the periods mentioned in the data, such as "Mojopahit," and the annotators often mixed up the role *style period* with the role *technique*—for example, in "Chiaroscuro." Although the number of training and evaluation examples for these roles was low, this suggests, first, that the annotators were not experienced enough to make consistent annotations on these roles and, second, that the vocabularies didn't always conform to the users' opinions.

Discussion

We wanted to investigate two aspects of the proposed automatic linguistic annotation method in this study: the accuracy our method could achieve in role identification compared to human annotators and the performance effect of using structured vocabularies as background knowledge. The results showed that the proposed method closely matched human performance and that performance improved using background knowledge. However, our method's performance differed in some roles. For example, the human

annotations for *creator*, *subject people*, and *title* were considerably more accurate than the automatic ones. A possible explanation could be that the sentence context wasn't sufficient to distinguish between the depicted and factual information about the persons. In addition, the vocabularies used often lacked the corresponding concepts for these roles.

We carried out the experiment with nonexpert annotators in a relatively specialized domain. The relatively low concordance among the annotators regarding the roles of *title*, *style period*, *cultural context*, and *technique* suggests that future research might compare the concordance of expert annotators with a data set and subsequently measure the performance of the method when more consistent training data is available.

Recent research in natural language processing and information extraction, such as statistical syntactic parsers and named-entity recognition systems,[10] have enabled advances in computational natural language understanding.[7] However, our study shows that a hybrid approach, using both statistical methods and rich background knowledge, results in higher performance. Of course, this is restricted to domains for which structured vocabularies are available. Previous research has achieved high accuracy in role identification when using hand-corrected parse trees on artificial data sets.[7,12] Nevertheless, it has been shown that these techniques generalize to other more closed domains only when appropriate training data is available.[14] This suggests that the performance of both the statistical tools used for the linguistic analysis and the vocabularies are dependent on the domain in which they're applied. Yet, the annotation method we propose is based on a feature set that could be applied to a domain similar to cultural heritage. For example, audio and video objects in the news domain use a similar metadata scheme and thus could apply our approach.

Because we concentrated on role identification in this study, we used a relatively simple method for concept identification. Although we obtained high accuracy in empirical evaluation for role identification, we didn't investigate the bias in concept identification. The full annotation could therefore require more sophisticated methods for concept identification.

With Web 2.0 user participation, the mass of knowledgeable amateurs could be used to further improve automatic annotation.

The study results revealed other areas for future work. The study didn't use information about a dynamic context,[12] which would address how other sentence constituents were classified. We only used features extracted from a single sentence and paths to its main verb. Adding features that would consider a more extensive context, rather than a single sentence, could lead to improved results.[15] Advanced classification strategies could also result in a performance gain.[12] For example, we could use separate classifiers to distinguish depiction information from factual information. We might improve performance with respect to the named entities by using anaphora or co-reference resolution. Vocabulary-based features different from the constituent vocabulary base and the vocabulary-root features are another area for exploration.

The automation of the annotation process is a key element in providing continuous access to digital cultural heritage. Moreover, with Web 2.0 user participation, the mass of knowledgeable amateurs together with cultural heritage professionals could be used to further improve automatic annotation. The required techniques to fully support automatic annotation of digital collections might still be debated, but we suspect hybrid approaches using both statistical and background-knowledge-based reasoning are required. This collective effort, mediated by machines and semantics, offers a promising way to annotate the ever-increasing volume of digital content.

Acknowledgments

We conducted this research as a part of MultimediaN e-Culture project (http://e-culture.multimedian.nl). We thank the Research Foundation of TKK for financial support, Michiel Hildebrand and Jan Wielenmaker for providing vocabulary services, and Laura Hollink for discussions on the evaluation methodology.

References

1. L. Hollink et al., "Classification of User Image Descriptions," *Int'l J. Human Computer Studies*, vol. 61, no. 5, 2004, pp. 501–626.
2. J. Kekäläinen and K. Järvelin, "The Co-effects of Query Structure and Expansion on Retrieval Performance in Probabilistic Text Retrieval," *Information Retrieval*, vol. 1, no. 4, 2000, pp. 329–344.
3. K-P. Yee et al., "Faceted Metadata for Image Search and Browsing, *Proc. SIGCHI Conf. Human Factors in Computing Systems*, ACM Press, 2003, pp. 401–408.
4. G. Schreiber et al., "Semantic Annotation and Search of Cultural-Heritage Collections: The MultimediaN e-Culture Demonstrator," *J. Web Semantics*, vol. 6, no. 4, 2008, pp. 243–249.
5. A.T. Schreiber et al., "Ontology-Based Photo Annotation," *IEEE Intelligent*

Systems, vol. 16, no. 3, 2001, pp. 66–74.
6. P. Buitelaar and T. Declerck, "Linguistic Annotation for the Semantic Web," *Annotation for the Semantic Web*, S. Handschuh and S. Staab, eds., IOS Press, 2003.
7. D. Gildea and D. Jurafsky, "Automatic Labeling of Semantic Roles," *Computational Linguistics*, vol. 28, no. 3, 2002, pp. 245–288.
8. D. Klein and C.D. Manning, "Fast Exact Inference with a Factored Model for Natural Language Parsing," *Advances in Neural Information Processing Systems* (NIPS 02), S. Becker, S. Thrun, and K. Obermayer, eds., MIT Press, 2002, pp. 3–10.
9. M. Marcus, B. Santorini, and M.A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, 1993, pp. 313–330.
10. J.R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," *Proc. 43rd Ann. Meeting Assoc. for Computational Linguistics* (ACL 05), Assoc. for Computational Linguistics, 2005, pp. 363–370.
11. E.F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," *Proc. 7th Conf. Natural Language Learning* (CoNLL 03), Morgan Kaufmann, 2003, pp. 142–147.
12. S. Pradhan et al., "Support Vector Learning for Semantic Argument Classification," *Machine Learning*, vol. 60, nos. 1–3, 2005, pp. 11–39.
13. A. Ben-David, "About the Relationship between ROC Curves and Cohen's Kappa," *Eng. Applications of Artificial Intelligence*, vol. 21, no. 6, 2008, pp. 874–882.
14. S.S. Pradhan, W. Ward, and J.H. Martin, "Towards Robust Semantic Role Labeling," *Computational Linguistics*, vol. 34, no. 2, 2008, pp. 289–310.
15. N. Xue and M. Palmer, "Calibrating Features for Semantic Role Labeling," *Proc. Conf. Empirical Methods in Natural Language Processing* (EMNLP 04), Assoc. of Computational Linguistics, 2004, pp. 88–94.

THE AUTHORS

Tuukka Ruotsalo is a PhD student in the Semantic Computing Research Group at the Helsinki University of Technology's Department of Media Technology. He also works as a researcher in the University of Helsinki's Department of Computer Science. His research interests include knowledge- and ontology-based methods for search, recommendation, and annotation of media content. Ruotsalo has an MSc in information systems science from the University of Jyväskylä. Contact him at tuukka.ruotsalo@cc.huttkk.fi.

Lora Aroyo is an assistant professor at the Vrije Universiteit Amsterdam. She leads the Cultural Heritage Information Presentation project with the Rijksmuseum Amsterdam on personalized museum tours and co-developed the IFanzy demonstrator for personalized TV program recommendation. She is technical coordinator of the EU's NoTube project on the integration of TV and Web data with the help of semantics. Her research interests include adaptive information management, recommendation systems, and ontology-based user and context modeling. Aroyo has a PhD in educational science and technology from the University of Twente. Contact her at l.m.aroyo@cs.vu.nl.

Guus Schreiber is a professor of intelligent information systems in Vrije Universiteit Amsterdam's Department of Computer Science and project coordinator of the EU's NoTube project on the integration of TV and Web data with the help of semantics.. His research interests are mainly in knowledge and ontology engineering, with a special interest in applications to cultural heritage. Schreiber has a PhD in social science informatics from the University of Amsterdam. Contact him at schreiber@cs.vu.nl.